Queueing Syst (2009) 62: 279–308
DOI 10.1007/s11134-009-9131-z

A polling model with an autonomous server

Roland de Haan · Richard J. Boucherie · Jan-Kees van Ommeren

Received: 4 October 2007 / Revised: 6 July 2009 / Published online: 30 July 2009

Abstract This paper considers polling systems with an autonomous server that remains at a queue for an exponential amount of time before moving to a next queue incurring a generally distributed switch-over time. The server remains at a queue until the exponential visit time expires, also when the queue becomes empty. If the queue is not empty when the visit time expires, service is preempted upon server departure, and repeated when the server returns to the queue. The paper first presents a necessary and sufficient condition for stability, and subsequently analyzes the joint queue-length distribution via an embedded Markov chain approach. As the autonomous exponential visit times may seem to result in a system that closely resembles a system of independent queues, we explicitly investigate the approximation of our system via a system of independent vacation queues. This approximation is accurate for short visit times only.

Keywords Single-server multi-queue system · Time-limited service discipline · Preemptive service · Unreliable server model

Mathematics Subject Classification (2000) 60K25 · 60K37

1 Introduction

Polling models are multi-queue systems with a single server. Typically, the server visits a queue, offers service to (a part of) the customers present at this queue, and then moves to a next queue. The specific details of the system may lead to quite distinct polling models. Polling models are typically characterized by: (i) the arrival process of the customers to the system (Poisson or more general), (ii) the service requirements

R. de Haan (✉) · R.J. Boucherie · J.-K. van Ommeren
University of Twente, Enschede, The Netherlands
e-mail: haanr@ewi.utwente.nl

of the customers, (iii) the servicing policy of the server (exhaustive, gated, k-limited, etc.), (iv) the visit order of the server, and (v) the switch-over times of the server between visits to the queues. Excellent surveys on a broad class of polling models can be found in, e.g., [1–3]. Applications of polling models are ubiquitous. For instance, traffic light systems, multiple-access protocols for communication networks (e.g., IEEE 802.11) and product-assembly systems can be modeled as a polling model.

In most of the (applications of) polling models, the server is assumed controllable. The goal is then to limit the time a server spends idle at a queue while there is still work in the system. To the contrary, in this article we assume that the server behaves autonomously (and thus is uncontrollable). More precisely, we assume that the server spends an exponentially distributed period of time at a queue irrespective of the number of customers present at each queue. A consequence of the autonomous server is that the services are subject to preemption. Applications of such specific polling models arise for instance in the context of wireless ad hoc networks in which cars, pedestrians or other moving objects that carry wireless equipment are used as communication hops.

The class of polling models that is most closely related to our model is the class of so-called *time-limited polling models* [4–7]. Leung [4] analyzes a time-limited model in which the server remains an exponential time at a queue but service is non-preemptive. Preemption is considered for a deterministic time-limited model by De Souza e Silva et al. [6] for Poisson arrivals and by Frigui and Alfa [5] for Markovian Arrival Processes. Eliazar and Yechiali [8, 9] studied a model with an exponential time limit and preemptive service. Observing that upon successful service completion at a queue the busy period in fact regenerates, the authors could obtain a direct, closed-form relation between the joint queue length at the end and the start of a server visit. In each of these models, the server is *impatient* and leaves a queue as soon as it becomes empty. A specific application of such a time-limited model to a timed token protocol can be found in [7].

A common assumption in polling model analysis is that the server moves to a next queue once the queue becomes empty. However, there also exists analytical work on models with a server that remains at a queue even when it becomes empty. These models are often referred to as patient server or stopping server models. The works of Eisenberg [10] and Borst [11] analyze several strategies for the server once the complete system becomes empty as to optimize some system performance measure. More recently, Boxma et al. [12, 13] considered a single-queue vacation model and a two-queue polling model in which the server upon arriving at an empty queue waits patiently for a certain duration before leaving again. We note that in the latter two-queue polling model (contrary to the models in [10] and [11]) there is no notion of work conservation anymore, since the server may wait patiently at one queue while the other queue is nonempty.

The only work we know of that includes both a given (random) visit time and a patient server that does not leave before the end of the visit time is [14]. This work considers the workload process for the autonomous server model with deterministic visit times. Due to the deterministic nature of the model, the queue lengths at the different queues can be decoupled and each queue is modeled as an M/G/1 queue with server vacations. Using an approximate analysis, the mean workload and mean message delay are studied.

For the case of a single queue, the polling model that we will consider boils down to the unreliable server model (USM) [15]. The extension of the analysis to a two-queue polling model appears feasible when the approach of, e.g., [16] or [17] would be followed. This approach requires to solve a boundary value problem. Unfortunately, this solution method appears an extremely difficult task for the two-queue model already, while for three or more queues analytical solutions in this direction are not anticipated.

In the first part of this article, we study a single-server polling model with $M \geq 1$ stations with infinite buffer in a stable environment. The main characteristics of the model are that the server visits a queue for a random amount of time (irrespective of the number of customers present at a queue) and that the service strategy is preemptive-repeat with resampling. Our interest is in the joint queue-length distribution at various instants in time. We should emphasize that the queue-length processes at the different queues in the system are not independent. For instance, the number of arrivals to the queues depends on the realizations of the random visit times at the other queues. Therefore, even for zero switch-over times, the queue-length processes are definitely not independent. However, we note that if the interest would only be in mean performance measures, then the queues could be considered in isolation. This is due to the fact that the vacation times at each of the queues are independent of the state of the system. Our analytical approach builds on the work of Eisenberg [18]. We set up a system of equations which relates the queue-length distributions at specific instants. The solution of this system is obtained via the explicit determination of the distribution at visit completion instants using an iterative approach. This approach is similar to the one introduced by Leung for probabilistically-limited polling models [19]. In the second part of this article, we study to what extent the queues in the polling system are independent. To this end, we propose an approximation based on the single-queue analysis of the USM. Next, we perform several numerical experiments to compare the results obtained from the polling model with the results based on the USM. In this way, we identify the system parameters for which the independence assumption on the individual queue lengths for the multi-queue system is justified.

This article is organized as follows. In Sect. 2, we describe the polling model. The analyses for the single-queue model and multi-queue model are given in Sects. 3 and 4, respectively. In Sect. 5, we study an approximation for the multi-queue model. The article is concluded in Sect. 6.

2 Model

Consider a set of M queues which are served by a single server. We denote queue i by Q_i, $i = 1, \ldots, M$ and references to Q_{i-1} or Q_{i+1} are always modulo M, e.g., for $i = 1$, Q_{i-1} refers in fact to Q_M. We will throughout use the subscript i to refer to a queue and for convenience leave out its range ($i = 1, \ldots, M$) whenever this does not lead to ambiguity. Customers arrive to Q_i according to a Poisson process with arrival rate λ_i. We denote the interarrival-time of customers at Q_i by I_i with distribution $I_i(t)$ and Laplace–Stieltjes Transform (LST) $\tilde{I}_i(s) = \lambda_i/(\lambda_i + s)$. A customer arriving

to Q_i requires an amount of service X_i with a distribution function $X_i(t)$, LST $\tilde{X}_i(s)$, and mean $1/\mu_i$. The service times are assumed to be independent.

A single server serves the queues at unit rate. For ease of presentation, we assume a fixed cyclic visit schedule $Q_1, Q_2, \ldots, Q_M, Q_1$, etc., but assuming other fixed cyclic schedules (e.g., in which queues are visited multiple times per cycle) would not significantly change the analysis. The server visits Q_i for an exponential amount of time denoted by Y_i with distribution $Y_i(t)$ and LST $\tilde{Y}_i(s) = \xi_i/(\xi_i + s)$. These visit times are assumed to be independent. The server always remains at a queue until the random visit time ends, even when the queue becomes empty. In other words, the dynamics of the server are not governed by the current state of the system. We assume that switch-over times C_i which occur when the server moves from Q_{i-1} to Q_i follow a distribution $C_i(t)$, with LST $\tilde{C}_i(s)$, and mean c_i. The switch-over times are assumed to be independent. Due to the patient nature of the server, (possibly multiple) idle periods can occur during a visit to a queue. The duration of each of these periods is distributed as the interarrival time at that queue.

We assume that customers are served according to the First-In-First-Out discipline. The service (but also the idle periods) at a queue are preempted at the end of a visit of the server. If there are some customers present at the beginning of the next visit, then a service time will be redrawn from the original distribution; thus, we adopt the so-called *preemptive-repeat with resampling* strategy.

The random processes C_i, I_i, X_i and Y_i are assumed to be independent.

3 Analysis of the single-queue model

The single-queue model comprises a single queue, say Q_1, fed by a Poisson process and with exponential visit times of the server. The intervisit times consist of the sum of the visit times to the other queues plus the total switch-over times. Our interest is in the marginal queue-length distribution at each queue of the polling system. From this distribution, performance measures such as the mean sojourn time of a customer may readily be derived. This single-queue model is in fact an unreliable server model or, alternatively, it may be considered as a vacation model with preemptive service. The first to analyze this specific model was Gaver [15] by introducing high priority customers (i.e., interruptions) and low priority customers (i.e., common arrivals). Here, we describe the unreliable server model, present the expression for the queue-length distribution and give a direct proof of this expression (which is somewhat more insightful than the proof in [15]).

Consider a sequence of alternating *processing* and *non-processing periods*. During a processing period, there are some customers at the queue and one of these is being served. During a non-processing period, no customers are present. The server may break down (and thus needs repair) at random points in time both during processing and non-processing periods. The *repair periods* have a duration D which follows a distribution $D(t)$ with LST $\tilde{D}(s)$ and mean $\mathbb{E}[D]$, and correspond to the intervisit times at Q_1 in our polling system. Thus, within the setting of the polling system, we have that $D = C_1 + \sum_{i \neq 1}(C_i + Y_i)$. We note that in this section, since only a single queue is considered, we will drop the subscript 1 whenever this does not lead to ambiguity. The periods between consecutive repairs, the so-called *availability periods*,

are assumed exponentially distributed with mean $1/\xi$ and these periods correspond to the visit times in the polling model. Customers arrive at the system according to a Poisson process with rate λ. We assume further that a preemptive-repeat servicing strategy with resampling is followed, i.e., if a service is interrupted, then at the start of the next availability period the service requirement is redrawn from the original service-time distribution.

We let $\tilde{X}_G(s)$ and $\mathbb{E}[X_G]$ denote the LST and the mean of the *generalized service time* of a customer, respectively. The latter period of time is defined as the period which starts when a customer receives service for the first time and ends when the customer leaves the system. We denote by $\hat{U}(z)$ the probability generating function (p.g.f.) of the number of customers that arrive during the service time of a customer which arrives to an empty system. Such a latter service (customer) will be referred to as an *exceptional first service* (*customer*). This service is exceptional in the sense that it may include the residual repair time of the server. Finally, let $\mathbb{E}[K]$ refer to the mean number of customers served during a processing period and define ρ_G as the generalized load of the system. Notice that it follows from our model assumptions that repair periods, availability periods and service times are independent.

Let us denote this queue-length distribution of the number of customers left behind by a departing customer by d_n, $n = 0, 1, 2, \ldots$. Then, the probability generating function $P_{L_d}(z)$ of the queue-length distribution is known (see, e.g., [15]) and given by the following theorem.

Theorem 1

$$P_{L_d}(z) = \frac{1}{\mathbb{E}[K]} \cdot \frac{\tilde{X}_G(\lambda(1-z)) - z \cdot \hat{U}(z)}{\tilde{X}_G(\lambda(1-z)) - z}, \tag{1}$$

where

$$\tilde{X}_G(s) = \frac{\tilde{X}(\xi+s) \cdot (\xi+s)}{(\xi+s) - \xi \cdot (1 - \tilde{X}(\xi+s)) \cdot \tilde{D}(s)},$$

$$\hat{U}(z) = \tilde{X}_G(\lambda(1-z)) \cdot \frac{\lambda z + \xi \cdot (\tilde{D}(\lambda(1-z)) - \tilde{D}(\lambda))}{z \cdot (\lambda + \xi(1 - \tilde{D}(\lambda)))},$$

$$\mathbb{E}[K] = \frac{1}{1-\rho_G} \cdot \frac{\lambda(1 + \xi \cdot \mathbb{E}[D])}{\lambda + \xi \cdot (1 - \tilde{D}(\lambda))}.$$

Notice that this departure distribution equals the time-equilibrium queue-length distribution. This can be argued by first using an up-and-down crossings argument and next appealing to the well-known Poisson Arrivals See Time Averages (PASTA) property [20]. Next, we will present several lemmas and defer the proof of the theorem until the end of this section.

Let us denote by V^* the processing time given that the service is interrupted. Further, we denote by X^* the service time given that the service is successful. Let V_j^* be i.i.d. copies of V^*, D_j be i.i.d. copies of D, and N a random variable denoting the

number of interruptions during a service. Notice that D, X^*, V_j^* and N are independent random variables. Then, the generalized service time X_G satisfies

$$X_G = X^* + \sum_{j=1}^{N} (V_j^* + D_j).$$

Lemma 1

$$\tilde{X}_G(s) = \mathbb{E}\big[e^{-sX_G}\big] = \frac{\tilde{X}(\xi+s)\cdot(\xi+s)}{(\xi+s) - \xi(1-\tilde{X}(\xi+s))\tilde{D}(s)}.$$

Proof The random variable N is geometrically distributed with success probability $\tilde{X}(\xi)$. The result for $\tilde{X}_G(s)$ follows by conditioning on N and some elementary calculus. □

The service time U of an exceptional first customer is given by

$$U = X_G + R_D \cdot \mathbf{1}_{\{R_D\}}, \tag{2}$$

where R_D denotes the residual repair time as seen by the first customer which arrives during a repair period, and $\mathbf{1}_{\{R_D\}}$ is the indicator function of the event that a customer that arrives during a repair time arrives to an empty system. It should be noted that X_G and R_D are independent. Let us introduce $N(T)$ to refer to the number of arrivals to the queue during a random period T, so that we can present the following lemma.

Lemma 2

$$\hat{U}(z) = \mathbb{E}\big[z^{N(U)}\big] = \mathbb{E}\big[z^{N(X_G)}\big] \cdot \mathbb{E}\big[z^{N(R_D\mathbf{1}_{\{R_D\}})}\big],$$

where

$$\mathbb{E}\big[z^{N(X_G)}\big] = \tilde{X}_G\big(\lambda(1-z)\big),$$

$$\mathbb{E}\big[z^{N(R_D\mathbf{1}_{\{R_D\}})}\big] = 1 - \big(1 - \mathbb{E}\big[z^{N(R_D)}\big]\big) \cdot \frac{\xi\cdot(1-\tilde{D}(\lambda))}{(\lambda+\xi) - \xi\cdot\tilde{D}(\lambda)}.$$

Proof By (2) and using that Poisson arrivals in disjoint intervals are independent, we have:

$$\hat{U}(z) = \mathbb{E}\big[z^{N(U)}\big] = \mathbb{E}\big[z^{N(X_G)+N(R_D\mathbf{1}_{\{R_D\}})}\big] = \mathbb{E}\big[z^{N(X_G)}\big] \cdot \mathbb{E}\big[z^{N(R_D\mathbf{1}_{\{R_D\}})}\big]. \tag{3}$$

Also, due to the Poisson assumption, we have:

$$\mathbb{E}\big[z^{N(X_G)}\big] = \tilde{X}_G\big(\lambda(1-z)\big).$$

Let us denote by $\mathbb{P}(\text{XFS})$ the probability that an arbitrary arriving customer has an exceptional first service (XFS). Then, we can write:

$$\begin{aligned}\mathbb{E}\big[z^{N(R_D\mathbf{1}_{\{R_D\}})}\big] &= \mathbb{E}\big[z^{N(R_D)}\big]\cdot\mathbb{P}(\text{XFS}) + 1\cdot\big(1-\mathbb{P}(\text{XFS})\big)\\ &= 1-\big(1-\mathbb{E}\big[z^{N(R_D)}\big]\big)\cdot\mathbb{P}(\text{XFS}).\end{aligned}$$

The p.g.f. $\mathbb{E}[z^{N(R_D)}]$ can be found by conditioning on the event of at least one arrival during the repair time:

$$\mathbb{E}\big[z^{N(R_D)}\big] = \frac{\mathbb{E}[z^{N(D)} \mid N(D)\geq 1]}{z} = \frac{\tilde{D}(\lambda(1-z))-\tilde{D}(\lambda)}{z(1-\tilde{D}(\lambda))}.$$

The probability $\mathbb{P}(\text{XFS})$ is obtained by considering its counterpart $\mathbb{P}(\overline{\text{XFS}}) = 1-\mathbb{P}(\text{XFS})$. The sequence of instants at which the queue becomes empty forms a renewal process. Note that the queue becomes empty only during an availability period and that the residual availability period is again exponentially distributed. Thus, by considering the first customer arriving after a renewal epoch, we can write a recursive relation for $\mathbb{P}(\overline{\text{XFS}})$:

$$\begin{aligned}\mathbb{P}\big(\overline{\text{XFS}}\big) &= \mathbb{P}\big(\{\text{arrival in availability period}\}\big)\\ &\quad+\big(1-\mathbb{P}\big(\{\text{arrival in availability period}\}\big)\big)\\ &\quad\times\mathbb{P}\big(\{\text{no arrival in the following repair period}\}\big)\cdot\mathbb{P}\big(\overline{\text{XFS}}\big).\end{aligned}$$

It follows that:

$$\mathbb{P}\big(\overline{\text{XFS}}\big) = \frac{\lambda}{\lambda+\xi} + \frac{\xi}{\lambda+\xi}\cdot\tilde{D}(\lambda)\cdot\mathbb{P}\big(\overline{\text{XFS}}\big) = \frac{\lambda}{\lambda+\xi\cdot(1-\tilde{D}(\lambda))}.$$

□

Proof of Theorem 1 Equation (1) can readily be obtained by studying the imbedded Markov chain at service completion instants (see, e.g., [21]). The explicit expressions that show up were derived in Lemmas 1 and 2. Finally, the term $\mathbb{E}[K]$ follows by inserting $z=1$ into (1) and applying L'Hospital's rule, yielding:

$$\mathbb{E}[K] = \frac{1}{1-\rho_G}\cdot\frac{\lambda(1+\xi\mathbb{E}[D])}{\lambda+\xi(1-\tilde{D}(\lambda))},$$

where

$$\rho_G = \lambda\cdot\mathbb{E}[X_G] = -\lambda\cdot\tilde{X}_G'(0).$$

□

4 Analysis of the multi-queue model

We have shown that for the polling system under consideration, the marginal queue-length distributions can be obtained by analyzing each queue in isolation. However,

the joint queue-length distribution cannot be obtained in this way due to the stochastics in the visit times of the server. Our analysis of the multi-queue model builds on the work of Eisenberg [18] which considers a polling model with a non-patient server and non-preemptive service. For this model, the queue-length distribution is determined at visit beginning, visit completion, service beginning, and service completion instants by studying the imbedded Markov chains defined at these instants. The fundamental relation in the analysis is the relation that counts the number of events with state $\mathbf{n}$ that occurred until time t [18, (4)]. In our work, we extend this relation for the polling model under consideration and we use this as a building block for obtaining the queue-length distribution at various instants.

We will first discuss the stability conditions of the system in Sect. 4.1. Next, in Sect. 4.2, we treat the extended counting relation in more detail. This counting relation is not sufficient to determine the queue-length distribution at all instants. To this end, we derive additional relations between the random variables in Sect. 4.3. However, even with these additional relations we still do not have enough information to solve our model completely. We will resolve this problem by deriving an explicit expression for the queue-length distribution at visit completion instants (see Sect. 4.4). This latter approach is based on work of Leung [19] for a probabilistically-limited polling model. Finally, we present the steady-state probabilities for the multi-queue model in Sect. 4.5.

4.1 Stability

The polling system is stable if and only if there exists a stationary regime in which each customer in the system can be served in a finite period of time. Due to the autonomous server, the vacation times for each individual queue are independent of the system state. Hence, the queues in the system can be considered in isolation with regard to stability. It follows that the system is stable if and only if all the queues in the system are stable, since service capacity cannot be exchanged among the queues. Thus, we can rely on the well-known non-saturation condition for the stability of a single (vacation) queue.

A necessary and sufficient condition for the stability of the polling system with the server operating under the autonomous-server discipline is given in the following theorem.

Theorem 2 (Autonomous-server discipline)

$$\textit{System is stable} \quad \Longleftrightarrow \quad \rho_i < \zeta_i, \quad \forall_{i\in\{1,\ldots,M\}},$$

where

$$\rho_i = \lambda_i \cdot \frac{1-\tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)},$$

$$\zeta_i = \frac{1/\xi_i}{\sum_{j=1}^{M}(1/\xi_j + c_j)},$$

where c_j is the mean switch-over time from Q_{j-1} to Q_j.

Proof It is well known that for a single queue the non-saturation condition is both a necessary and sufficient condition for stability, i.e.,

$$Q_i \text{ is stable} \quad \Longleftrightarrow \quad \rho_i < \zeta_i, \quad i = 1, \ldots, M,$$

where ρ_i is the mean effective amount of work arriving per time unit to Q_i and ζ_i is the availability fraction of the server at Q_i.

Consider first the mean effective amount of work arriving per time unit to Q_i. This amount is determined by the total number of customers arriving per time unit λ_i and the mean effective amount of work each individual brings for the server, denoted by $\mathbb{E}[X_i^{\text{eff}}]$, as follows:

$$\rho_i = \lambda_i \cdot \mathbb{E}\big[X_i^{\text{eff}}\big].$$

The quantity $\mathbb{E}[X_i^{\text{eff}}]$ is in fact the mean total time the server spends on serving a customer at Q_i including any interrupted services. Noting that the number of interruptions per customer is geometrically distributed, it can be found via simple calculus that:

$$\mathbb{E}\big[X_i^{\text{eff}}\big] = \frac{1 - \tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)}.$$

The availability fraction of the server ζ_i is fully specified by the mean visit times, the visit frequencies and the switch-over times between the queues. Notice that a complete cycle consists of a_i visits to Q_i, $i = 1, \ldots, M$, and the switch-over times between the queues. It then readily follows for the availability fraction of the server at Q_i:

$$\zeta_i = \frac{1/\xi_i}{\sum_{j=1}^{M}(1/\xi_j + c_j)}.$$

It is good to notice that the fraction ζ_i is independent of the load at the queues. The observation that the system is stable if and only if all the queues in the system are stable completes the proof. □

4.2 A relation for the queue-length distribution at various instants

We set up a relation for the number of occurrences of specific events. Apart from the events defined in [18], we define a number of additional events. We introduce events related to the start and the completion of an idle period. These events do not appear in Eisenberg's model as in his model the server leaves a queue as soon as it becomes empty. Moreover, we introduce events related to the interruption of a service or idle period due to the end of a server visit. Let us denote by n_i the number of customers at Q_i. Next, we can define the following variables which all refer to the number of the given events with state $\mathbf{n} = (n_1, \ldots, n_M)$ that occur in $(0, t)$ at Q_i:

$\omega^i(t; \mathbf{n})$, service beginnings.

$\pi^i(t; \mathbf{n})$, successful service completions (i.e., server does not switch during service).

$\pi_*^i(t; \mathbf{n})$, interrupted services (i.e., server switches during service).

$\alpha^i(t;\mathbf{n})$, visit beginnings.

$\beta^i(t;\mathbf{n})$, visit completions.

$a^i(t;\mathbf{n})$, idle period beginnings.

$b^i(t;\mathbf{n})$, idle period completions (i.e., server does not switch during idle period).

$b^i_*(t;\mathbf{n})$, interrupted idle periods (i.e., server switches during idle period).

We note that $\mathbf{n}$ refers to the number of customers present in the system (either waiting or in service) immediately after the specific event occurred. These variables are related in the following way for all $\mathbf{n} \in \mathbb{N}^M$ and $t \geq 0$:

$$\left[\pi^i(t;\mathbf{n})+\pi^i_*(t;\mathbf{n})\right]+\alpha^i(t;\mathbf{n})+\left[b^i(t;\mathbf{n})+b^i_*(t;\mathbf{n})\right]=\omega^i(t;\mathbf{n})+\beta^i(t;\mathbf{n})+a^i(t;\mathbf{n}). \tag{4}$$

This counting relation should be as read as follows. At each instant that one of the events present at the l.h.s. of (4) with state $\mathbf{n}$ occurs, also exactly one event with the same state $\mathbf{n}$ at the r.h.s. occurs. For instance, a visit beginning event at Q_i at time t with state $\mathbf{n}'$, $\pi^i_*(t;\mathbf{n}')$, always coincides exactly either with a service beginning event at Q_i with the same state $\mathbf{n}'$, $\omega^i(t;\mathbf{n}')$ (if there are some customers present upon the server's arrival) or with an idle period beginning at Q_i with the same state $\mathbf{n}'$, $a^i(t;\mathbf{n}')$ (if no customers are present upon the server's arrival).

We note that the end of a server visit always corresponds to an interruption event and vice versa. Therefore, we can isolate these events and break up (4) into:

$$\pi^i_*(t;\mathbf{n})+b^i_*(t;\mathbf{n})=\beta^i(t;\mathbf{n}), \tag{5}$$

$$\pi^i(t;\mathbf{n})+\alpha^i(t;\mathbf{n})+b^i(t;\mathbf{n})=\omega^i(t;\mathbf{n})+a^i(t;\mathbf{n}). \tag{6}$$

Let us define imbedded Markov chains each corresponding to instants at which one of the counting processes increases. Each state in a Markov chain is uniquely defined by the position i of the server ($i=1,\ldots,M$) and $\mathbf{n}=(n_1,\ldots,n_M)$, the number of customers present in the system at a certain instant. We define the steady-state probabilities for each event type by dividing the number of events with state $\mathbf{n}$ that occurred until t by the total number of the events until t, and then taking the limit for t to infinity, yielding:

$$\alpha^i_{\mathbf{n}}=\lim_{t\to\infty}\left[\alpha^i(t;\mathbf{n})/\alpha^i(t)\right], \qquad \beta^i_{\mathbf{n}}=\lim_{t\to\infty}\left[\beta^i(t;\mathbf{n})/\beta^i(t)\right],$$
$$b^i_{\mathbf{n}}=\lim_{t\to\infty}\left[b^i(t;\mathbf{n})/b^i(t)\right], \qquad b^i_{*,\mathbf{n}}=\lim_{t\to\infty}\left[b^i_*(t;\mathbf{n})/b^i_*(t)\right],$$
$$a^i_{\mathbf{n}}=\lim_{t\to\infty}\left[a^i(t;\mathbf{n})/a^i(t)\right], \qquad \omega^i_{\mathbf{n}}=\lim_{t\to\infty}\left[\omega^i(t;\mathbf{n})/\omega(t)\right],$$
$$\pi^i_{\mathbf{n}}=\lim_{t\to\infty}\left[\pi^i(t;\mathbf{n})/\pi(t)\right], \qquad \pi^i_{*,\mathbf{n}}=\lim_{t\to\infty}\left[\pi^i_*(t;\mathbf{n})/\pi_*(t)\right],$$

where

$$\alpha^i(t) = \sum_{\mathbf{n}} \alpha^i(t;\mathbf{n}), \qquad \beta^i(t) = \sum_{\mathbf{n}} \beta^i(t;\mathbf{n}), \qquad b^i(t) = \sum_{\mathbf{n}} b^i(t;\mathbf{n}),$$
$$b^i_*(t) = \sum_{\mathbf{n}} b^i_*(t;\mathbf{n}), \qquad a^i(t) = \sum_{\mathbf{n}} a^i(t;\mathbf{n}), \qquad \omega(t) = \sum_i \sum_{\mathbf{n}} \omega^i(t;\mathbf{n}),$$
$$\pi(t) = \sum_i \sum_{\mathbf{n}} \pi^i(t;\mathbf{n}), \qquad \pi_*(t) = \sum_i \sum_{\mathbf{n}} \pi^i_*(t;\mathbf{n}).$$

It can be seen that for a stable system all these limits exist with probability one by using renewal theory arguments. For instance, consider $\pi^i_{\mathbf{n}}$ as given above. We have that for given i and $\mathbf{n}$, $\{\pi^i(t;\mathbf{n}),\ t \geq 0\}$ forms a renewal process, thus $\lim_{t\to\infty}[\pi^i(t;\mathbf{n})/t]$ exists with probability one (see, e.g., [22]). Moreover, it follows that under stability, $\lim_{t\to\infty}[\pi(t)/t]$ exists, and thus we can conclude that the ratio of these latter limits exists. The other probabilities can be argued analogously and are thus also correctly defined.

Notice that (hereby following [18]) we have that all probabilities are conditioned on Q_i except for $\omega^i_{\mathbf{n}}$, $\pi^i_{\mathbf{n}}$ and $\pi^i_{\mathbf{n},*}$. Along with the steady-state probabilities, let us also define the corresponding p.g.f.'s as follows:

$$\alpha^i(\mathbf{z}) = \sum_{\mathbf{n}} \alpha^i_{\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}}, \qquad \beta^i(\mathbf{z}) = \sum_{\mathbf{n}} \beta^i_{\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}}, \qquad b^i(\mathbf{z}) = \sum_{\mathbf{n}} b^i_{\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}},$$
$$b^i_*(\mathbf{z}) = \sum_{\mathbf{n}} b^i_{*,\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}}, \qquad a^i(\mathbf{z}) = \sum_{\mathbf{n}} a^i_{\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}}, \qquad \omega(\mathbf{z}) = \sum_{\mathbf{n}} \omega^i_{\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}},$$
$$\pi(\mathbf{z}) = \sum_{\mathbf{n}} \pi^i_{\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}}, \qquad \pi_*(\mathbf{z}) = \sum_{\mathbf{n}} \pi^i_{*,\mathbf{n}} \cdot \mathbf{z}^{\mathbf{n}},$$

where $\mathbf{z}^{\mathbf{n}} := z_1^{n_1} \cdots z_M^{n_M}$.

Next, we divide (5) and (6) by $\pi(t)$ and take the limit of $t \to \infty$, yielding:

$$\pi^i_{*,\mathbf{n}} \lim_{t\to\infty}\big[\pi_*(t)/\pi(t)\big] + b^i_{*,\mathbf{n}} \lim_{t\to\infty}\big[b^i_*(t)/\pi(t)\big] = \beta^i_{\mathbf{n}} \lim_{t\to\infty}\big[\beta^i(t)/\pi(t)\big], \tag{7}$$

$$\pi^i_{\mathbf{n}} \lim_{t\to\infty}\big[\pi(t)/\pi(t)\big] + \alpha^i_{\mathbf{n}} \lim_{t\to\infty}\big[\alpha^i(t)/\pi(t)\big] + b^i_{\mathbf{n}} \lim_{t\to\infty}\big[b^i(t)/\pi(t)\big]$$
$$= \omega^i_{\mathbf{n}} \lim_{t\to\infty}\big[\omega(t)/\pi(t)\big] + a^i_{\mathbf{n}} \lim_{t\to\infty}\big[a^i(t)/\pi(t)\big]. \tag{8}$$

Using similar renewal arguments as above, it is readily verified that under our model assumptions all these limits indeed exist with probability one if the stability conditions are satisfied.

Let us introduce some notation. We denote by $p_{pr,X}$ the probability of an arbitrary service at some queue in the system being preempted and by $p^i_{pr,I}$ the probability of an idle period at Q_i being preempted (i.e., the server switches before the next customer arrives to the queue). The mean cycle time of the server will be denoted by $\mathbb{E}[C]$. Finally, we define $\kappa_i := \lim_{t\to\infty}[a^i(t)/\pi(t)]$. This enables us to present the following theorem.

Theorem 3 *The p.g.f.'s of the queue-length distribution at Q_i, $i = 1, \ldots, M$, at specific imbedded instants in a polling model with an autonomous server are related as follows*:

$$\frac{p_{pr,X}}{1 - p_{pr,X}} \cdot \pi_*^i(z) + \kappa_i \cdot p_{pr,I}^i \cdot b_*^i(z) = \gamma \cdot \beta^i(z),$$

$$\pi^i(z) + \gamma \cdot \alpha^i(z) + \kappa_i \cdot \left(1 - p_{pr,I}^i\right) \cdot b^i(z) = \frac{\omega^i(z)}{1 - p_{pr,X}} + \kappa_i \cdot a^i(z),$$

where

$$p_{pr,X} = 1 - \frac{\sum_j \lambda_j}{\sum_j \lambda_j / \tilde{X}_j(\xi_j)},$$

$$p_{pr,I}^i = 1 - \tilde{I}_i(\xi_i), \quad i = 1, \ldots, M,$$

$$\kappa_i = \frac{1}{p_{pr,I}^i} \cdot \left(\gamma - \frac{\lambda_i}{\sum_j \lambda_j} \cdot \frac{1 - \tilde{X}_i(\xi_i)}{\tilde{X}_i(\xi_i)}\right), \quad i = 1, \ldots, M,$$

$$\gamma = \frac{1}{\sum_j \lambda_j \mathbb{E}[C]}.$$

It should be noted that $1/X_i(\xi_i)$ refers to the mean number of required server visits to serve a single customer. Next, we will present several lemmas and defer the proof of the theorem until the end of this section.

Lemma 3

$$\lim_{t\to\infty}\left[\alpha^i(t)/\pi(t)\right] = \lim_{t\to\infty}\left[\beta^i(t)/\pi(t)\right] = \frac{1}{\sum_j \lambda_j \mathbb{E}[C]}, \quad i = 1, \ldots, M.$$

Proof First, notice that the number of visit completions, $\beta^i(t)$, differs by at most one from the number of visit beginnings, $\alpha^i(t)$, for any $t \geq 0$. Therefore, we have that:

$$\lim_{t\to\infty}\left[\alpha^i(t)/\beta^i(t)\right] = 1.$$

Second, the number of visit beginnings at Q_i per cycle is exactly one. Hence, it follows that:

$$\lim_{t\to\infty}\left[\alpha^i(t)/t\right] = \frac{1}{\mathbb{E}[C]},$$

where for $\mathbb{E}[C]$, the mean cycle time, we have:

$$\mathbb{E}[C] = \sum_j \left(\frac{1}{\xi_j} + c_j\right). \tag{9}$$

Third, the average total number of service completions per cycle is equal to the average total number of arrivals per cycle (assuming a stable system). This implies that:

$$\lim_{t\to\infty}\bigl[\pi(t)/t\bigr] = \sum_j \lambda_j.$$

Combining these three limits yields the desired result. □

Lemma 4

$$\lim_{t\to\infty}\bigl[\omega(t)/\pi(t)\bigr] = \frac{1}{1-p_{pr,X}},$$
$$\lim_{t\to\infty}\bigl[\pi_*(t)/\pi(t)\bigr] = \frac{p_{pr,X}}{1-p_{pr,X}}.$$

Proof The $\lim_{t\to\infty}[\omega(t)/\pi(t)]$ is defined as the limit of the ratio of the total number of service beginnings and the total number of (successful) service completions. The numerator and denominator are related via the probability of an arbitrary service being preempted, $p_{pr,X}$. More precisely,

$$\lim_{t\to\infty}\bigl[\pi(t)/\omega(t)\bigr] = 1 - p_{pr,X}.$$

Similarly to the relation between $\alpha^i(t)$ and $\beta^i(t)$, we note that $\omega(t)$ and $\pi(t) + \pi_*(t)$ differ at most one for $t \geq 0$. Therefore, we can write:

$$\lim_{t\to\infty}\bigl[\pi_*(t)/\pi(t)\bigr] = \lim_{t\to\infty}\bigl[\bigl(\omega(t)-\pi(t)\bigr)/\pi(t)\bigr] = \frac{p_{pr,X}}{1-p_{pr,X}}.$$

□

Lemma 5

$$\lim_{t\to\infty}\bigl[b^i(t)/\pi(t)\bigr] = \kappa_i\cdot\bigl(1-p^i_{pr,I}\bigr), \quad i=1,\ldots,M,$$
$$\lim_{t\to\infty}\bigl[b^i_*(t)/\pi(t)\bigr] = \kappa_i\cdot p^i_{pr,I}, \quad i=1,\ldots,M.$$

Proof Recall that we set $\lim_{t\to\infty}[a^i(t)/\pi(t)] =: \kappa_i$, where κ_i is a constant yet to be determined. These limits do not have a simple interpretation, but we can relate them to limits for other events. The number of events $a^i(t)$ and $b^i(t)$ are related as follows:

$$\lim_{t\to\infty}\bigl[b^i(t)/a^i(t)\bigr] = 1 - p^i_{pr,I},$$

where $p^i_{pr,I}$, the probability that an idle period at Q_i is preempted, depends on i, and is given by:

$$p^i_{pr,I} = 1 - \tilde{I}_i(\xi_i).$$

Analogously, $a^i(t)$ and $b^i_*(t)$ are related via:

$$\lim_{t\to\infty}\left[b^i_*(t)/a^i(t)\right] = p^i_{pr,I},$$

which completes the proof. □

Proof of Theorem 3 The presented equations follow by first evaluating the limit expressions in (7) and (8). The limit expressions are derived in the lemmas above. However, these expressions still contain the unknowns $p_{pr,X}$ and $\kappa_i,\ i = 1,\ldots,M$.

For the service preemption probability $p_{pr,X}$, we obtain:

$$\begin{aligned} p_{pr,X} &= \sum_j \mathbb{P}(\text{service is preempted} \mid \text{s.b. at } Q_j)\cdot\mathbb{P}(\text{s.b. at } Q_j \mid \text{s.b. at some queue}) \\ &= \sum_j \left(1-\tilde{X}_j(\xi_j)\right)\cdot\mathbb{P}(\text{s.b. at } Q_j \mid \text{s.b. at some queue}) \\ &= \frac{\sum_j (1-\tilde{X}_j(\xi_j))\lambda_j/\tilde{X}_j(\xi_j)}{\sum_k \lambda_k/\tilde{X}_k(\xi_k)} = 1 - \frac{\sum_j \lambda_j}{\sum_k \lambda_k/\tilde{X}_k(\xi_k)}, \end{aligned}$$

where we use that:

$$\begin{aligned} &\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue}) \\ &\quad= \frac{\lambda_i/(1-\mathbb{P}(\text{serv. at } Q_i \text{ is preempted} \mid \text{s.b. at } Q_i))}{\sum_j \lambda_j/(1-\mathbb{P}(\text{serv. at } Q_j \text{ is preempted} \mid \text{s.b. at } Q_j))} \\ &\quad= \frac{\lambda_i/\tilde{X}_i(\xi_i)}{\sum_j \lambda_j/\tilde{X}_j(\xi_j)}, \end{aligned} \tag{10}$$

where the condition *s.b. at some queue* in fact means that we consider the imbedded Markov chain of all service beginning instants. Notice that multiple service beginning events may correspond to a single customer.

The unknown $\kappa_i,\ i = 1,\ldots,M$, can be found from (8) (or alternatively from (7)) by inserting all the limit expressions and summing both sides over $\mathbf{n}$. After several rearrangements and using that

$$\sum_{\mathbf{n}} \pi^i_{*,\mathbf{n}} = \mathbb{P}(\text{s.i. at } Q_i \mid \text{s.i. at some queue}) = \frac{\lambda_i/\tilde{X}_i(\xi_i) - \lambda_i}{\sum_j (\lambda_j/\tilde{X}_j(\xi_j) - \lambda_j)}, \tag{11}$$

where we use *s.i.* as short for service interruption, we eventually obtain:

$$\kappa_i = \frac{1}{p^i_{pr,I}}\left(\gamma - \frac{\lambda_i}{\sum_j \lambda_j}\cdot\frac{1-\tilde{X}_i(\xi_i)}{\tilde{X}_i(\xi_i)}\right).$$

The final step is to write these equations in terms of p.g.f.'s by multiplication with $\mathbf{z}^{\mathbf{n}}$ and summation over $\mathbf{n}$. □

4.3 Additional relations for the queue-length distributions at specific instants

We need additional relations to obtain the queue-length distributions at the different instants defined. Eisenberg [18] presents relations between $\pi^i(\mathbf{z})$ and $\omega^i(\mathbf{z})$ for the non-patient server model with non-preemptive services. We show that with a minor modification this relation can be used to relate both $\pi^i(\mathbf{z})$ and $\omega^i(\mathbf{z})$ and $\pi^i_*(\mathbf{z})$ and $\omega^i(\mathbf{z})$ in our model. Moreover, relations between $a^i(\mathbf{z})$ and $b^i(\mathbf{z})$ and between $a^i(\mathbf{z})$ and $b^i_*(\mathbf{z})$ can be established in a similar fashion. Finally, for completeness we repeat the relation from [18] between $\alpha^i(\mathbf{z})$ and $\beta^{i-1}(\mathbf{z})$.

Recall that $\omega^i(\mathbf{z})$, $\pi^i_*(\mathbf{z})$ and $\pi^i(\mathbf{z})$ refer to the number of customers at all queues at instants of service beginning, service interruption and successful service completion, respectively. The relations between these quantities are given in the following lemma.

Lemma 6

$$\begin{aligned}
\pi_i(z) &= \frac{\tilde{X}_i(\xi_i)\cdot\left(\sum_j \lambda_j/\tilde{X}_j(\xi_j)\right)}{\sum_j \lambda_j}\cdot \breve{X}_i(z)\cdot\frac{\omega_i(z)}{z_i},\\
\pi^i_*(z) &= \left(1-\tilde{X}_i(\xi_i)\right)\cdot\left(\sum_j \lambda_j/\tilde{X}_j(\xi_j)-\lambda_j\right)\cdot X^*_i(z)\cdot\omega_i(z),
\end{aligned} \tag{12}$$

where

$$\begin{aligned}
\breve{X}_i(z) &= \frac{\tilde{X}_i\left(\xi_i+\sum_j \lambda_j(1-z_j)\right)}{\tilde{X}_i(\xi_i)},\\
X^*_i(z) &= \frac{\xi_i}{\xi_i+\sum_j \lambda_j(1-z_j)}\cdot\frac{1-\tilde{X}_i\left(\xi_i+\sum_j \lambda_j(1-z_j)\right)}{1-\tilde{X}_i(\xi_i)}.
\end{aligned}$$

Proof Let us first consider (12), i.e., the relation between $\omega^i(\mathbf{z})$ and $\pi^i(\mathbf{z})$. We note that every successful service completion instant has a corresponding service beginning instant, while the correspondence the other way round is not true due to preemption (which is caused by the exogenously determined visit times of the server). Notice further that the fact whether a service will get interrupted does not depend on the queue-length distribution at the start of a service.

Unfortunately, we cannot relate $\omega^i(\mathbf{z})$ and $\pi^i(\mathbf{z})$ in the straightforward manner as was done by Eisenberg [18]. This is due to the preemption assumption which no longer ensures that the long-term fraction of all service beginnings that occur at Q_i and the long-term fraction of all service completions that occur at Q_i are equal for all parameter settings. Or, in mathematical terms, the relation $\omega^i(\mathbf{z})\,|_{\mathbf{z}=\mathbf{1}} = \pi^i(\mathbf{z})\,|_{\mathbf{z}=\mathbf{1}}$, $\forall_i$, is no longer necessarily true.

Recall first the definitions of $\omega^i(\mathbf{z})$ and $\pi^i(\mathbf{z})$:

$$\omega^i(\mathbf{z}) = \sum_{n_1}\cdots\sum_{n_M} z_1^{n_1}\cdots z_M^{n_M}\,\mathbb{P}\big(\{\mathbf{N}=\mathbf{n}\}\cap\{\text{s.b. at } Q_i\}\mid \text{s.b. at some queue}\big),$$

$$\pi^i(\mathbf{z}) = \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \mathbb{P}\big(\{\mathbf{N} = \mathbf{n}\} \cap \{\text{s.c. at } Q_i\} \mid \text{s.c. at some queue}\big),$$

where *s.c.* is used as short for service completion. Then, to circumvent the use of these unconditional p.g.f.'s, we define $\omega_c^i(\mathbf{z})$ and $\pi_c^i(\mathbf{z})$ as follows:

$$\omega_c^i(\mathbf{z}) := \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \cdot \mathbb{P}(\mathbf{N} = \mathbf{n} \mid \text{s.b. at } Q_i),$$

$$\pi_c^i(\mathbf{z}) := \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \cdot \mathbb{P}(\mathbf{N} = \mathbf{n} \mid \text{s.c. at } Q_i),$$

where

$$\mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue}) = \frac{\lambda_i}{\sum_j \lambda_j}. \tag{13}$$

This latter equation follows immediately by the observation that the number of arriving customers per time unit is equal to the number of served customers per time unit for a system in equilibrium. According to the definitions above, we have that:

$$\omega^i(\mathbf{z}) = \omega_c^i(\mathbf{z}) \cdot \mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue}),$$

$$\pi^i(\mathbf{z}) = \pi_c^i(\mathbf{z}) \cdot \mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue}).$$

Notice that $\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})$ was already given in (10).

Using that the distribution of the number of customers at the start of a service is independent of the success of the service, we can relate the conditional p.g.f.'s in the following manner:

$$\pi_c^i(\mathbf{z}) = \frac{\breve{X}_i(\mathbf{z})}{z_i} \cdot \omega_c^i(\mathbf{z}), \tag{14}$$

where the term $1/z_i$ is due to the fact that the number of customers at Q_i at a service completion instant is exactly one less than at the service beginning instant and $\breve{X}_i(\mathbf{z})$ is the p.g.f. of the number of customers that arrive at all queues during a service at Q_i that is indeed completed. The latter is given by:

$$\breve{X}_i(\mathbf{z}) := \mathbb{E}\big[\mathbf{z}^{N(X_i)} \mid X_i < Y_i\big] = \frac{\mathbb{E}[\mathbf{z}^{N(X_i)} \mathbf{1}_{\{X_i < Y_i\}}]}{\mathbb{P}(X_i < Y_i)} = \frac{\tilde{X}_i(\xi_i + \sum_j \lambda_j (1 - z_j))}{\tilde{X}_i(\xi_i)}, \tag{15}$$

where we used the notation $N(T)$ to refer to the number of arrivals to all queues during a random period T. The final equality sign follows from first conditioning on X_i and Y_i and next using that $\mathbb{E}[\mathbf{z}^{N(x)}]$ is Poisson distributed with parameter $\sum_j \lambda_j \cdot (1 - z_j) \cdot x$ for a given x. Combining the definitions of the conditional p.g.f.'s and (14), we obtain:

$$\pi_i(\mathbf{z}) = \frac{\mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue})}{\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})} \cdot \breve{X}_i(\mathbf{z}) \cdot \frac{\omega_i(\mathbf{z})}{z_i}. \tag{16}$$

The relation between $\pi_*^i(\mathbf{z})$ and $\omega_i(\mathbf{z})$ is derived analogously and it resembles (16):

$$\pi_*^i(\mathbf{z}) = \frac{\mathbb{P}(\text{s.i. at } Q_i \mid \text{s.i. at some queue})}{\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})} \cdot X_i^*(\mathbf{z}) \cdot \omega_i(\mathbf{z}), \tag{17}$$

where

$$X_i^*(\mathbf{z}) := \mathbb{E}\big[\mathbf{z}^{N(Y_i)} \mid X_i > Y_i\big] = \frac{\xi_i}{\xi_i + \sum_j \lambda_j(1-z_j)} \cdot \frac{1 - \tilde{X}_i(\xi_i + \sum_j \lambda_j(1-z_j))}{1 - \tilde{X}_i(\xi_i)}.$$

The derivation of $X_i^*(\mathbf{z})$ is done analogously to the derivation of $\breve{X}_i(\mathbf{z})$. Notice that the term $1/z_i$ is absent in (17), since no customer departs from the queue. The final step of the proof is to substitute the conditional probabilities of (10), (11) and (13) into (16) and (17). □

Remark 1 We note that for non-preemptive service the first ratio on the r.h.s. of (16) equals one as a service beginning corresponds uniquely to a service completion. Further, in this case, we have that the term $\breve{X}_i(\mathbf{z})$ equals $\mathbb{E}[z^{N(X_i)}]$, so that we obtain (17) of [18].

Recall that $a^i(\mathbf{z})$, $b_*^i(\mathbf{z})$ and $b^i(\mathbf{z})$ refer to the number of customers at instants of idle period beginning, idle period interruption and idle period completion at Q_i, respectively. The relations between these quantities are given in the following lemma.

Lemma 7

$$b^i(z) = \breve{I}_i(z) \cdot z_i \cdot a^i(z),$$
$$b_*^i(z) = \breve{I}_i(z) \cdot a^i(z),$$

where

$$\breve{I}_i(z) = \frac{\tilde{I}_i(\xi_i + \sum_{j\neq i} \lambda_j(1-z_j))}{\tilde{I}_i(\xi_i)}.$$

Proof Let us first consider the relation between $a^i(\mathbf{z})$ and $b^i(\mathbf{z})$. We note that every idle period completion instant has a corresponding idle period beginning instant, while the correspondence the other way round is not true. This is due to the exponential visit time of the server. Whether the idle period gets interrupted depends on the arrival process and on the distribution of the visit time of the server only. In particular, it does not depend on the queue-length distribution at the start of an idle period. Thus, we can relate the generating functions $a^i(\mathbf{z})$ and $b^i(\mathbf{z})$ by the following observations. The p.g.f. of the number of customers that arrive at all queues different from Q_i during an idle period that it is completed with an arrival is given by:

$$\breve{I}_i(\mathbf{z}) := \mathbb{E}\big[\mathbf{z}^{N(I_i)} \mid I_i < Y_i\big] = \frac{\tilde{I}_i(\xi_i + \sum_{j\neq i} \lambda_j(1-z_j))}{\tilde{I}_i(\xi_i)}.$$

This expression can be derived in a similar fashion as (15). Further, we note that exactly one customer arrives at Q_i at the end of the idle period. Together, this yields the following relation between $a_i(\mathbf{z})$ and $b_i(\mathbf{z})$:

$$b^i(\mathbf{z}) = \breve{I}_i(\mathbf{z}) \cdot z_i \cdot a^i(\mathbf{z}).$$

In the same manner, the relation between $b^i_*(\mathbf{z})$ and $a^i(\mathbf{z})$ can be established:

$$b^i_*(\mathbf{z}) = \breve{I}_i(\mathbf{z}) \cdot a^i(\mathbf{z}).$$

We note that $\mathbb{E}[\mathbf{z}^{N(Y_i)} \mid I_i > Y_i] = \mathbb{E}[\mathbf{z}^{N(I_i)} \mid I_i < Y_i] =: \breve{I}_i(\mathbf{z})$, since both Y_i and I_i are assumed exponentially distributed. □

Recall that $\alpha^i(\mathbf{z})$ and $\beta^i(\mathbf{z})$ refer to the number of customers at visit beginning instants and visit completion instants at Q_i, respectively. The relations between these quantities are given in the following lemma.

Lemma 8

$$\alpha^i(z) = \hat{C}_i(z) \cdot \beta^{i-1}(z), \tag{18}$$

where,

$$\hat{C}_i(z) = \tilde{C}_i\left(\sum_j \lambda_j(1 - z_j)\right).$$

Proof There exists a well-known relation (see, e.g., [18]) between the number of customers that the server leaves behind in the system at departure from Q_{i-1} and the number of customers in the system that the server finds upon arrival to Q_i. This difference is characterized by the number of arriving customers during a switch-over time from Q_{i-1} to Q_i. We denote by $\hat{C}_i(\mathbf{z})$ the p.g.f. of this number, which is given by:

$$\hat{C}_i(\mathbf{z}) = \tilde{C}_i\left(\sum_j \lambda_j(1 - z_j)\right),$$

since arrivals to the queues are according to a Poisson process. Hence, we immediately obtain (18). □

Altogether, we have derived $7 \cdot M$ relations between the $8 \cdot M$ p.g.f.'s of our interest. By combining the equations from Lemmas 6, 7 and 8 with the ones of Theorem 3, we are able to fully specify all the p.g.f.'s in terms of $\beta^i(\mathbf{z})$. It can then be shown that it is sufficient to determine the M p.g.f.'s for $\beta^i(\mathbf{z})$, $i = 1, \ldots, M$, explicitly; the latter will be done below.

4.4 Queue-length probabilities at visit completion instants via auxiliary variables

We shall determine the p.g.f. of the queue-length distribution at visit completion instants, $\beta^i(\mathbf{z})$, explicitly. This part of the analysis is based on work by Leung [19] for the study of a probabilistically-limited polling model, which was later extended in [4] to a time-limited polling model, and involves setting up an iterative scheme. A key role in this iterative scheme is played by the (auxiliary) p.g.f.'s $\phi_k(\mathbf{z})$ and $\phi_k^s(\mathbf{z})$, which will be explained below. In the final step of the iteration scheme, $\beta^i(\mathbf{z})$ is obtained as a simple function of $\phi_k^s(\mathbf{z})$.

We consider a tagged queue i and we will leave out the subscript and superscript i whenever it does not lead to ambiguity. Let again $N(T)$ denote the number of arrivals to all queues during a random period T while $\mathbf{1}_{\{A\}}$ denotes the indicator function for event A. The explicit expression for $\beta^i(\mathbf{z})$ is stated in the following lemma.

Lemma 9

$$\beta^i(z) = \sum_{k=1}^{\infty} \phi_k^s(z),$$

where

$$\begin{aligned}
\phi_k^s(z) &= \phi_{k-1}(z)|_{z_i=0} \cdot \left(\mathbb{E}\left[z^{N(Y)}\mathbf{1}_{\{Y<I\}}\right] + z_i \mathbb{E}\left[z^{N(I)}\mathbf{1}_{\{Y>I\}}\right] \cdot \mathbb{E}\left[z^{N(Y)}\mathbf{1}_{\{Y<X\}}\right]\right) \\
&\quad + \left(\phi_{k-1}(z) - \phi_{k-1}(z)|_{z_i=0}\right) \cdot \mathbb{E}\left[z^{N(Y)}\mathbf{1}_{\{Y<X\}}\right], \\
\phi_k(z) &= \phi_{k-1}(z)|_{z_i=0} \cdot \left(z_i \mathbb{E}\left[z^{N(I)}\mathbf{1}_{\{Y>I\}}\right] \cdot \mathbb{E}\left[z^{N(X)}\mathbf{1}_{\{Y>X\}}\right] \cdot \frac{1}{z_i}\right) \\
&\quad + \left(\phi_{k-1}(z) - \phi_{k-1}(z)|_{z_i=0}\right) \cdot \mathbb{E}\left[z^{N(X)}\mathbf{1}_{\{Y>X\}}\right] \cdot \frac{1}{z_i}.
\end{aligned}$$

Proof Let us introduce first the concept of a *service period*. A service period is defined as a period which starts either at a visit beginning or at a service completion instant, and which ends with either the next service completion instant or an interruption (due to the departure of the server), whichever occurs first. It is important to notice that the final service period of a server's visit always ends with an interruption (and thus comprises no successful service completion), while each of the other service periods comprises exactly one successfully completed service. Also notice that the first service period always starts at a visit beginning instant. Let us denote by $\phi_k(\mathbf{z})$, $k \geq 1$, the p.g.f. of the number of customers at all queues at the end of the kth service period and service period k is not the final service period (i.e., service period k ends with a successful service completion, and service period $k+1$ will occur). Similarly, we denote by $\phi_k^s(\mathbf{z})$, $k \geq 1$, the number of customers at all queues at the end of the kth service period and k is in fact the final service period (i.e., service period k will be interrupted, and service period $k+1$ will not occur). Finally, we define $\phi_0(\mathbf{z}) := \alpha(\mathbf{z})$, i.e., $\phi_0(\mathbf{z})$ is the p.g.f. of the number of customers present at the start of a visit. Then, $\phi_k^s(\mathbf{z})$ and $\phi_k(\mathbf{z})$, $k = 1, 2, \ldots$, are given by:

$$\phi_k^s(\mathbf{z}) = \phi_{k-1}(\mathbf{z})|_{z_i=0} \cdot \left(\mathbb{E}\left[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<I\}}\right] + z_i \mathbb{E}\left[\mathbf{z}^{N(I)}\mathbf{1}_{\{Y>I\}}\right] \cdot \mathbb{E}\left[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X\}}\right]\right)$$

$$+\left(\phi_{k-1}(\mathbf{z})-\phi_{k-1}(\mathbf{z})|_{z_i=0}\right)\cdot\mathbb{E}\left[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X\}}\right], \tag{19}$$

and

$$\phi_k(\mathbf{z})=\phi_{k-1}(\mathbf{z})|_{z_i=0}\cdot\left(z_i\mathbb{E}\left[\mathbf{z}^{N(I)}\mathbf{1}_{\{Y>I\}}\right]\cdot\mathbb{E}\left[\mathbf{z}^{N(X)}\mathbf{1}_{\{Y>X\}}\right]\frac{1}{z_i}\right)$$
$$+\left(\phi_{k-1}(\mathbf{z})-\phi_{k-1}(\mathbf{z})|_{z_i=0}\right)\cdot\mathbb{E}\left[\mathbf{z}^{N(X)}\mathbf{1}_{\{Y>X\}}\right]\frac{1}{z_i}, \tag{20}$$

where

$$\mathbb{E}\left[\mathbf{z}^{N(I)}\mathbf{1}_{\{Y>I\}}\right]=\tilde{I}_i\left(\xi_i+\sum_{j\neq i}\lambda_j(1-z_j)\right), \tag{21}$$

$$\mathbb{E}\left[\mathbf{z}^{N(X)}\mathbf{1}_{\{Y>X\}}\right]=\tilde{X}_i\left(\xi_i+\sum_{j}\lambda_j(1-z_j)\right), \tag{22}$$

$$\mathbb{E}\left[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<I\}}\right]=\frac{\xi_i}{\xi_i+\sum_{j\neq i}\lambda_j(1-z_j)}\cdot\left(1-\tilde{I}_i\left(\xi_i+\sum_{j\neq i}\lambda_j(1-z_j)\right)\right), \tag{23}$$

$$\mathbb{E}\left[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X\}}\right]=\frac{\xi_i}{\xi_i+\sum_{j}\lambda_j(1-z_j)}\cdot\left(1-\tilde{X}_i\left(\xi_i+\sum_{j}\lambda_j(1-z_j)\right)\right). \tag{24}$$

Let us elaborate on (19). Recall that for $\phi_k^s(\mathbf{z})$ a service period k is indeed the final (i.e., interrupted) service period. Clearly, the number of customers at all queues at the end of service period k is equal to the number present at the end of the service period $k-1$, plus the ones that arrived during the present service period. The length of service period k, and thus the number of arriving customers, depends on whether one or more customers were present at the end of the previous service period, which explains why the right-hand side of the equation consists of two parts. The first part refers to the case that at the end of the previous service period no customers were present at Q_i. In this case, we must also distinguish between whether an interruption occurred during the idle period or during the possibly associated service. The second part refers to the case that some customers were present at Q_i at the end of the previous service period. Equation (20) can be derived analogously. The expressions in (21), (22), (23) and (24) follow from simple probabilistic calculations. It is also worthwhile to note that $\phi_0(\mathbf{1})=1$, while $\phi_k(\mathbf{1})<1$, for all $k=1,2,\ldots$, since the $(k+1)$-st service period need not occur at all during a visit to Q_i. Finally, observing that there is a one-to-one relationship between a visit completion and the end of a final service period, we can write

$$\beta^i(\mathbf{z})=\sum_{k=1}^{\infty}\phi_k^s(\mathbf{z}),$$

which completes the proof. □

Algorithm 1 Pseudo-code of the iterative scheme for determining $\check{\beta}^i(\mathbf{k}), \forall_i, \forall_{\mathbf{k}}$

$\check{\beta}^{i_0}(\mathbf{k}) = 1,\ \forall_{i_0},\ \forall_{\mathbf{k}}$; (*start with an empty system*)
FOR $i_1 = 1, \ldots, M$
  set $i_2 := i_1$;
  REPEAT
    $\check{\beta}_r^{i_2}(\mathbf{k}) = \check{\beta}^{i_2}(\mathbf{k}),\ \forall_{\mathbf{k}}$;
    set $j := 0$;
    set $\check{\phi}_0(\mathbf{k}) = \check{\beta}^{i_2-1}(\mathbf{k}) \cdot \check{C}_{i_2}(\mathbf{k})$;
    REPEAT
      set $j := j + 1$;
      compute $\check{\phi}_j(\mathbf{k}),\ \forall_{\mathbf{k}}$, *using (20)*;
      compute $\check{\phi}_j^s(\mathbf{k}),\ \forall_{\mathbf{k}}$, *using (19)*;
      compute $\check{\beta}^{i_2}(\mathbf{k}) = \sum_{l=1}^{j} \check{\phi}_l^s(\mathbf{k}),\ \forall_{\mathbf{k}}$;
    UNTIL $1 - \mathrm{Re}(\check{\beta}^{i_2}(\mathbf{0})) < \delta$
    set $i_2 := \mathrm{MOD}(i_2, M) + 1$;
  UNTIL $|\mathrm{Re}(\check{\beta}^{i_1}(\mathbf{k})) - \mathrm{Re}(\check{\beta}_r^{i_1}(\mathbf{k}))| < \epsilon,\ \forall_{\mathbf{k}}$
END

We set up an iterative scheme to compute $\beta^i(\mathbf{z})$ numerically. The scheme is constructed in terms of Discrete Fourier Transforms (DFTs) as these appear more convenient for computational purposes. To this end, we replace $z_i, \forall_i$, in the expressions above by $\omega_i^{k_i}$, where $\omega_i = \exp(-2\pi I/H_i)$, so that all expressions become functions of $\mathbf{k} = (k_1, \ldots, k_M)$. Here I is the imaginary unit and H_i refers to the number of points used for Q_i to determine the joint probabilities. The DFT of a variable is denoted by adding the accent, $\check{}$, to the variable, e.g., we let $\check{\beta}^i(\mathbf{k})$ refer to the DFT of $\beta^i(\mathbf{z})$. The pseudo-code of the iterative scheme is presented in Algorithm 1. The standard values for the convergence parameters that have been used are $\epsilon = 10^{-6}$ and $\delta = 10^{-9}$. Finally, via the Inverse Fourier Transform, the steady-state probabilities $\beta_{\mathbf{n}}^i$ are obtained. These probabilities are only exact for $H_i \to \infty,\ i = 1, \ldots, M$, but the strength of the approach is that in general the probabilities are already close to the exact values for small values of H_i. However, it should be noted that when the system load increases, these values H_i must typically be increased to guarantee the accurate computation of the probabilities. Thus, this iterative approach appears mainly applicable to systems with a light to moderate load.

Remark 2 The p.g.f. $\pi^i(z)$, which refers to the queue-length at service completion instants, can now be obtained using the derived relations (see Sects. 4.2, 4.3) and the explicit computation of $\beta^i(z)$. However, $\pi^i(z)$ can also directly be expressed in terms of the introduced auxiliary p.g.f. $\phi_k(z)$ as follows:

$$\pi^i(z) = \frac{\mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c.})}{\mathbb{E}[\#\text{ s.c. per visit to } Q_i]} \cdot \sum_{k=1}^{\infty} \phi_k(z).$$

Remark 3 In our model, interruptions can occur both during services and during idle periods, while in Leung's time-limited model (see [4]) only services can be interrupted. The latter is due to the fact that in Leung's model the server moves to the next queue if there are no customers present anymore. Due to the additional event of idle period interruptions in our model, the probability that $\psi_i(j) \geq 1$ (one or more customers present at Q_i after j services) of (9) of [19] which is conditioned on the event that no interruption occurs during the jth service is no longer equal to the unconditional probability. Nevertheless, we strongly believe that for our model the approach of [19] could still be followed to find $\beta^i(z)$. However, the expressions will become quite involved, so that we proposed an unconditional approach here.

4.5 Steady-state queue-length probabilities and sojourn times

The exponential visit times allow us to obtain the steady-state queue-length probabilities. To see this, consider the full queue-length process and condition on the server being at some tagged queue Q. Next, remove from this full process the time periods that the server is not at Q and concatenate the remaining parts. This induced process then consists of a series of exponentially distributed periods with jumps between each two periods. These jumps in fact constitute a Poisson batch arrival process and reflect the arrivals to the system when the server is not at Q. Due to the PASTA property, these batches see time average behavior upon arrival. Moreover, the system observed by the arriving batches is exactly the system as observed by the server when it departs from Q. Thus, we have that a departing server observes the system in steady-state conditioned on the position of the server.

Let us denote the p.g.f. of the number of customers present at a random instant during a switch-over time from Q_{i-1} to Q_i by $\hat{C}_i^R(\mathbf{z})$. It is well known that this p.g.f. should satisfy:

$$\hat{C}_i^R(\mathbf{z}) = \beta^i(\mathbf{z}) \cdot \frac{1 - \tilde{C}_i(\sum_j \lambda_j(1 - z_j))}{c_i \cdot (\sum_j \lambda_j(1 - z_j))}.$$

Hence, by conditioning on the position of the server (either it is visiting a queue or switching between two queues), we may write for $P(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}}]$, the joint p.g.f. of the steady-state queue lengths,

$$P(\mathbf{z}) = \frac{1}{\mathbb{E}[C]} \cdot \sum_{i=1}^{M} \left(\beta^i(\mathbf{z}) \cdot \frac{1}{\xi_i} + \hat{C}_i^R(\mathbf{z}) \cdot c_i \right), \tag{25}$$

where $\mathbb{E}[C]$ is given in (9).

It should be noted that in the discussion above the size of the arriving batches depends on the realizations of the random visit times at the other queues. Therefore, even for zero switch-over times, the queue-length processes at the different queues are definitely not independent. Besides, it is good to notice that the acquirement of these probabilities is not a very common result in the polling literature. For most polling models that have been analyzed, no steady-state queue-length probabilities are known.

Let us next turn to the marginal distribution. We denote by $P_i(z)$ the marginal queue-length distribution for Q_i, which readily follows from $P(\mathbf{z})$, i.e.,

$$P_i(z) = P(z_1, \ldots, z_M)|_{z_j=1,\ j\neq i,\ z_i=z}.$$

The marginal probabilities can also be obtained via $\pi^i(\mathbf{z})$ (see Remark 2). Using an up-and-down crossings argument and the PASTA property [20], we may write:

$$P_i(z) = \frac{\pi^i(z_1, \ldots, z_M)|_{z_j=1,\ j\neq i,\ z_i=z}}{\pi^i(1, \ldots, 1)}.$$

We note that this relation does not rely on the exponentiality of the visit times and it may also be applied for the analysis of other service disciplines. Alternatively, the marginal distribution $P_i(z)$ could be obtained using the techniques discussed in Sect. 3. This is due to the fact that the random vacation times at each of the queues are independent of the state of the system.

From the marginal queue-length distribution, the LST of the sojourn time (or delay) of a customer, which we denote by $\tilde{D}_i(s)$, can be obtained using the distributional form of Little's law (see [23]). In particular, we have that:

$$\tilde{D}_i(s) = P_i(z)|_{z=1-s/\lambda_i}.$$

Thus, all moments of the sojourn time at each queue can be determined.

Remark 4 (Single-queue model) For the special case $M = 1$, the expression for $P(z)$, (25), is known in the literature (see also (1)). However, the equivalence is not apparent. Notice that our method presented in this section provides a numerical scheme that yields identical numerical values, though it does not give an explicit formula. This may seem a drawback, but our method is designed for $M \geq 2$.

5 Approximations

We have investigated the correlations among the queue lengths in the polling model for a wide range of parameter settings. The outcomes of this investigation have led to a proposed approximation and corresponding performance measure. This will all be discussed in Sect. 5.1. Next, in Sect. 5.2, we perform a numerical study on the quality of the approximation.

5.1 Dependence of the queue lengths

5.1.1 Correlations

As an example, we present results for a symmetric system with three queues, exponential service times and zero switch-over times. For ease of presentation, we define

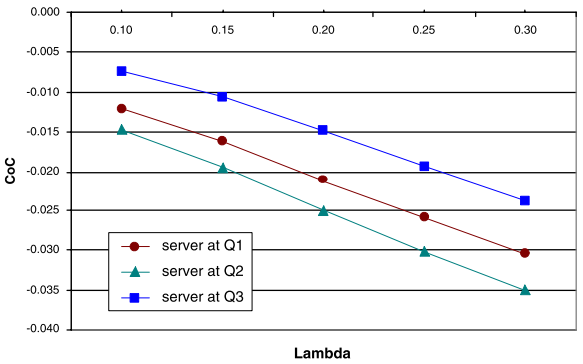

Fig. 1 The coefficient of correlation (CoC) as a function of Λ for $\mu = 1.00$ and $\xi = 1.00$ (exponential service times)

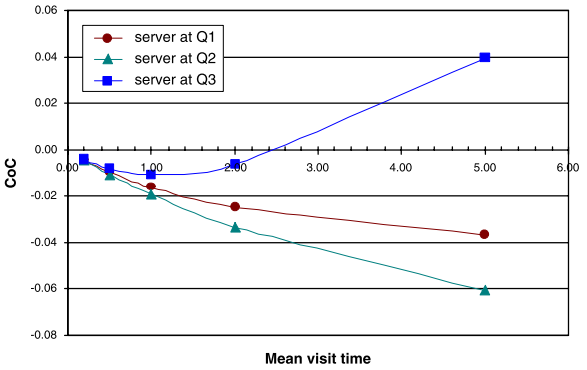

Fig. 2 The coefficient of correlation as a function of the mean visit time $(1/\xi)$ for $\Lambda = 0.15$ and $\mu = 1.00$ (exponential service times)

$\Lambda = \sum_j \lambda_j$, $\mu = \mu_i$, and $\xi = \xi_i$, for $i = 1, \ldots, M$. Specifically, we consider the coefficient of correlation, $\rho_{1,2|Q_j}$, $j = 1, 2, 3$, for the conditional queue length at Q_1 and Q_2 as a function of Λ and ξ, where $\rho_{1,2|Q_j}$ is defined as follows:

$$\rho_{1,2|Q_j} := \frac{\mathrm{Cov}(N_1\ , N_2 \mid \text{server at } Q_j)}{\sqrt{\mathrm{Var}(N_1 \mid \text{server at } Q_j)\mathrm{Var}(N_2 \mid \text{server at } Q_j)}}$$
$$= \frac{\mathbb{E}[N_1 \cdot N_2 \mid \text{server at } Q_j] - \mathbb{E}[N_1 \mid \text{server at } Q_j] \cdot \mathbb{E}[N_2 \mid \text{server at } Q_j]}{\sqrt{\mathrm{Var}(N_1 \mid \text{server at } Q_j)\mathrm{Var}(N_2 \mid \text{server at } Q_j)}}.$$

Notice that we only consider the conditional queue lengths here. This is because the system state generally depends heavily on the position of the server, so that it is more meaningful to compare conditional probabilities. Also, if we would take a snapshot of the system state at a random instant in time, then we do not expect it to be in line with the unconditional time-equilibrium probabilities.

In Fig. 1, we plot $\rho_{1,2|Q_j}$ as a function of the arrival rate Λ for the situation $\mu = 1.00$ and $\xi = 1.00$. It is shown that the correlation between the queues is quite small (for all server's positions), although it increases (in absolute sense) slightly in Λ. Figure 2 shows the impact of the mean visit time to a queue $(= 1/\xi)$ on $\rho_{1,2|Q_j}$ for the situation $\Lambda = 0.15$ and $\mu = 1.00$. The plot shows that the coefficient of correlation is small for short visit times, but that it may drift rapidly away when the visit times

grow large. This is in accordance with the fact that for $1/\xi \downarrow 0$ the queue lengths indeed become independent (under zero switch-over times), yielding a coefficient of correlation equal to zero. We have also generated results for many other parameter settings for the symmetric three-queue system. These results demonstrate that for a wide range of settings the coefficient of correlation is quite small which indicates little dependence between the queue lengths at the different queues.

5.1.2 Approximation

A natural next step is then to study approximations for the joint queue-length distribution of the polling model based on the assumption of independence of the queues. Such approximations could be of great value since our experiments have shown that the computation time for the joint queue-length probabilities in the polling model may grow quite large.

Our approximation for the joint queue-length distribution is thus based on the marginal distributions. These marginal distributions can be computed almost directly via the unreliable server model (USM) (see Sect. 3). We note that these single-queue results can be obtained very fast which is often a necessity for real applications. Specifically, the approximation reads as follows:

$$\mathbb{P}(N_1 = n_1, \ldots, N_M = n_M \mid \text{server at } Q_j) \approx \prod_{i=1}^{M} \mathbb{P}(N_i = n_i \mid \text{server at } Q_j). \quad (26)$$

To assess the quality of this approximation, we compute the terms on the r.h.s. of (26) via the USM. As we have not analyzed these terms yet, this will be done next.

Let us consider the unreliable server model with arrival rate λ, service rate μ, exponentially distributed availability periods with parameter ξ. We denote the Erlang$_{M-1}(\xi)$ distributed repair periods by D, and its LST by $\tilde{D}(\cdot)$. Individual (exponential) repair stages are denoted by D^j, $j = 1, \ldots, M-1$, with LST $\tilde{D}^j(\cdot)$. We let the p.g.f. $\hat{N}_{1j}(z) = \mathbb{E}[z^{N(Q_1)} | \text{server at } Q_j]$, $j = 1, \ldots, M$, refer to the number of customers in the queue given that the server is either at the queue (for $j = 1$) or at "stage" $j-1$ of the repair period (for $j \neq 1$). Notice that (due to exponentially distributed availability periods) $\hat{N}_{11}(z)$ in fact refers to the p.g.f. of the number of customers present at an arbitrary instant of the availability period. Denote further by $\hat{N}_{1D}(z)$ the p.g.f. of the number of customers present at an arbitrary instant of the repair period. These quantities are related to $P_{L_d}(z)$ (see (1)) as follows:

$$P_{L_d}(z) = p_a \hat{N}_{11}(z) + p_r \hat{N}_{1D}(z),$$

where p_a $(= 1/M)$ and p_r $(= 1 - p_a)$ are the long-term fractions that the server is available and being repaired, respectively. Observe that $\hat{N}_{11}(z)$ and $\hat{N}_{1D}(z)$ are also related via:

$$\hat{N}_{1D}(z) = \hat{N}_{11}(z) \cdot \hat{D}_A(z),$$

where $\hat{D}_A(z)$ is the p.g.f. of the number of arrivals from the start of the repair period until an arbitrary instant of that period, and satisfies, using simple regenerative

processes theory (see, e.g., [24]),

$$\hat{D}_A(z) = \frac{1 - \hat{D}(z)}{\hat{D}'(1)(1-z)},$$

where $\hat{D}(z)$ $(= \tilde{D}(\lambda(1-z)))$ is the p.g.f. of the number of arrivals during the complete repair period.

Hence, it follows that:

$$\hat{N}_{11}(z) = \frac{P_{L_d}(z)}{p_a + p_r \hat{D}_A(z)}.$$

We note that $\hat{N}_{1j}(z)$, $j \neq 1$, can be decomposed in three independent parts. The first part refers to the number of customers present at the end of an availability period. The second part accounts for the arrivals during the already completed repair stages. Finally, the last part represents the number of arrivals from the beginning of repair stage $j-1$ until a random instant during this stage. In terms of p.g.f.'s, this leads to:

$$\hat{N}_{1j}(z) = \hat{N}_{11}(z) \cdot \prod_{k=1}^{j-2} \hat{D}^k(z) \cdot \hat{D}_A^j(z), \quad j = 2, \ldots, M,$$

where $\hat{D}^k(z)$ refers to the arrivals during the (completed) kth stage of the repair period and is given by

$$\hat{D}^k(z) = \tilde{D}^k\big(\lambda(1-z)\big), \quad k = 1, \ldots, M-2,$$

and $\hat{D}_A^j(z)$ (cf. $\hat{D}_A(z)$) is given by

$$\hat{D}_A^j(z) = \frac{1 - \hat{D}^j(z)}{\hat{D}'^j(1)(1-z)}, \quad j = 2, \ldots, M.$$

Finally, the probabilities $\mathbb{P}(N_i = n_i | \text{server at } Q_j)$ are obtained from $\hat{N}_{1j}(z)$ using DFT techniques. Notice that for a comparison with an asymmetric polling system, all steps above have to be performed for each queue separately.

The proposed approximation is anticipated to work well in situations where the individual queues behave independently. In our polling model, it seems that under our imposed visit-time distribution the dependencies between the different queues are small (see also Sect. 5.1.1). For instance, the number of arrivals during the absence of the server and the time that a queue is served are known (in distribution) and independent of what occurs at the other queues in the system.

We have now all the tools at hand to investigate the dependencies between the queues in the polling system. Let us emphasize that our objective here is not to perform an exhaustive numerical study for all system parameters and service time distributions. The underlying idea of the approximation is that if the queues in the system would turn out to be "almost" independent, then the results of a much simpler single-queue model can be used as a good approximation for a complex multi-queue polling model. Therefore, our purpose is mainly to gain preliminary insight in the parameter ranges for which the approximation works well.

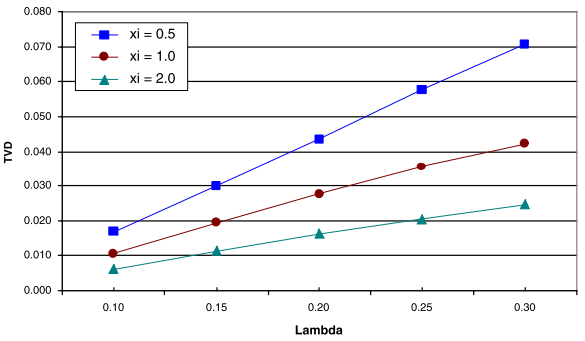

Fig. 3 The total variation distance (TVD) as a function of Λ (exponential service times)

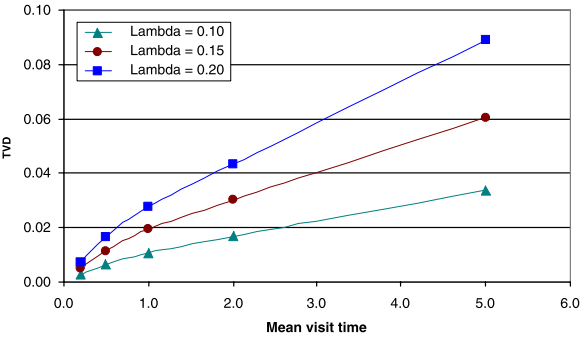

Fig. 4 The total variation distance (TVD) as a function of the mean visit time $(1/\xi)$ (exponential service times)

5.2 Numerical results

We present here numerical results from experiments for a symmetric three-queue polling model for both exponentially and deterministically distributed service times. The mean service time $1/\mu$ is set equal to one. The performance measure that we adopt to assess the quality of the approximation is as follows. We use the measure of total variation distance [25] for the queue-length distribution conditional on the position of the server, denoted by $\theta^p_{\text{cond},j}$:

$$\theta^p_{\text{cond},j} := \sum_{\mathbf{n}} \Big| \mathbb{P}(N_1 = n_1, \ldots, N_M = n_M \mid \text{server at } Q_j) - \prod_{i=1}^{M} \mathbb{P}(N_i = n_i \mid \text{server at } Q_j) \Big|.$$

For ease of presentation, we define $\theta^p_{\text{cond}} := \theta^p_{\text{cond},j}$, for $j = 1, \ldots, M$. This measure quantifies the difference between the exact and the approximate distribution. Clearly, if the approximation is exact (e.g., when the queue lengths are independent), θ^p_{cond} equals zero, and it is strictly positive otherwise.

The results for the total variation distance in the exponential case are presented in Figs. 3 and 4. First, consider Fig. 3 in which θ^p_{cond} is plotted as a function of Λ for

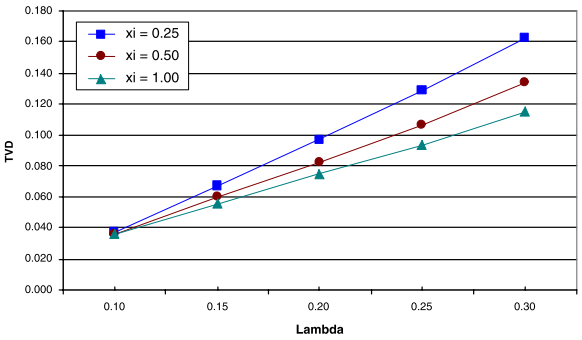

Fig. 5 The total variation distance (TVD) as a function of Λ (deterministic service times)

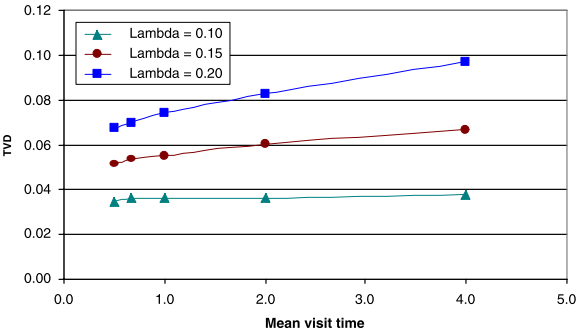

Fig. 6 The total variation distance (TVD) as a function of the mean visit time (deterministic service times)

various values of ξ. The slopes observed in this figure clearly show that θ^p_{cond} is not insensitive to Λ, but increases linearly in the arrival rate. Moreover, it can be seen that θ^p_{cond} decreases in ξ. This is further illuminated in Fig. 4 which shows the impact of the mean visit time (i.e., $1/\xi$) on θ^p_{cond} for various values of Λ. It is shown that θ^p_{cond} is quite small for short visit times and increases linearly in $1/\xi$ for longer visit times.

The results for deterministic service times are presented in Figs. 5 and 6. Figure 5 shows θ^p_{cond} as a function of Λ for various values of ξ. Again as for the exponential case, θ^p_{cond} increases linearly in Λ. The impact of ξ on θ^p_{cond} appears small. This is confirmed by the plot of Fig. 5 which shows the total variation distance as a function of the mean visit time for various values of Λ. However, an important difference with respect to the exponential case is that θ^p_{cond} stays away from zero when the mean visit time decreases. The latter is due to the fact that the load for the deterministic case increases in ξ, so that the queue lengths will not approach independence under the stable regime (i.e., $\rho_{G,i} < 1,\ \forall_i$).

Let us wrap up the main observations that we have done in our experiments for the three-queue symmetric system: (i) θ^p_{cond} is positively correlated to the arrival rate Λ; (ii) θ^p_{cond} decreases rapidly toward zero when the mean visit times are shortened for exponential service times, while for deterministic service θ^p_{cond} seems to decreases to an asymptotic value strictly larger than zero.

We have seen that there exists a wide range of parameter settings for which the approximation works quite well. However, the approximation appears not applicable

to heavily loaded systems. For such situations, it might be worthwhile to consider heavy-traffic approximations. This will be part of future work.

6 Conclusions

Polling models with an autonomous server may arise as a performance model in the context of mobile wireless technologies. We have analyzed this polling model in great detail by determining the queue-length distribution at specific instants. Due to the state-space expansion under heavy load, our analytical approach appears mainly applicable to systems with a light to moderate load. We have also performed several experiments to study the dependence between queue lengths. This has led to the identification of system parameter settings for which a simple single-queue vacation model can successfully be applied to approximate performance measures for the complex polling model. These experiments show that the quality of the approximation is not very sensitive to the total arrival rate, but mainly depends on the mean visit time. The shorter the visit times, the better will be the approximation for the polling model measures.

In future work, we will study other network structures such as a (multi-hop) chain model or a multi-path model. We strongly believe that similar techniques as described above may prove useful to analyze such models. Another meaningful extension is to include more complex mobility patterns and even models with multiple servers might be considered.

Acknowledgements We would like to thank the anonymous referees for their constructive comments. We also want to express our gratitude to Sindo Núñez-Queija for his extensive comments on an earlier draft of this article. This work was supported by Easy Wireless—Ministry of Economic Affairs, Department of Commerce, under Grant IS043014. In the Netherlands, the 3 Universities of Technology have formed the 3TU.Federation. This article is the result of joint research in the 3TU.Centre of Competence NIRICT (Netherlands Institute for Research on ICT).

References

1. Takagi, H.: Queueing analysis of polling systems: An update. In: Stochastic Analysis of Comput. Commun. Syst., pp. 267–318 (1990)
2. Takagi, H.: Queueing analysis of polling models: progress in 1990–1994. In: Dshalalow, J.H. (ed.) Frontiers in Queueing: Models, Methods and Problems, pp. 119–146. CRC Press, Boca Raton (1997)
3. Vishnevskii, V.M., Semenova, O.V.: Mathematical methods to study the polling systems. Automat. Remote Control **67**(2), 173–220 (2006)
4. Leung, K.K.: Cyclic-service systems with non-preemptive time-limited service. IEEE Trans. Commun. **42**(8), 2521–2524 (1994)
5. Frigui, I., Alfa, A.-S.: Analysis of a time-limited polling system. Comput. Commun. **21**(6), 558–571 (1998)
6. de Souza e Silva, E., Gail, H.R., Muntz, R.R.: Polling systems with server timeouts and their application to token passing networks. IEEE Trans. Netw. **3**(5), 560–575 (1995)

7. Tangemann, M., Sauer, K.: Performance analysis of the timed token protocol of FDDI and FDDI-II. IEEE J. Sel. Areas Commun. **9**(2), 271–278 (1991)
8. Eliazar, I., Yechiali, U.: Polling under the randomly timed gated regime. Stoch. Models **14**(1–2), 79–93 (1998)
9. Eliazar, I., Yechiali, U.: Randomly timed gated queueing systems. SIAM J. Appl. Math. **59**(2), 423–441 (1998)
10. Eisenberg, M.: Two queues with changeover times. Oper. Res. **19**(2), 386–401 (1971)
11. Borst, S.C.: A polling system with a dormant server. Report BS-R9313, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands (1993)
12. Boxma, O.J., Schlegel, S., Yechiali, U.: A note on an M/G/1 queue with a waiting server, timer and vacations. In: Analytic Methods in Applied Probability: In Memory of Fridrikh Karpelevich. American Mathematical Society Translations, vol. 2(207), pp. 25–35. AMS, Providence (2002)
13. Boxma, O.J., Schlegel, S., Yechiali, U.: Two-queue polling models with a patient server. Ann. Oper. Res. **112**, 101–121 (2002)
14. Xie, J., Fischer, M.J., Harris, C.M.: Workload and waiting time in a fixed-time loop system. Comput. Oper. Res. **24**(8), 789–803 (1997)
15. Gaver, D.P.: A waiting line with interrupted service, including priorities. J. R. Stat. Soc. **24**(1), 73–90 (1962)
16. Fayolle, G., Iasnogorodski, R.: Two coupled processors: The reduction to a Riemann–Hilbert problem. Z. Wahrscheinlichkeitstheor. Verw. Geb. **47**, 325–351 (1979)
17. Coffman, E.G., Fayolle, G., Mitrani, I.: Two queues with alternating service periods. In: Performance '87: Proc. of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation, pp. 227–239 (1988)
18. Eisenberg, M.: Queues with periodic service and changeover times. Oper. Res. **20**(2), 440–451 (1972)
19. Leung, K.K.: Cyclic-service systems with probabilistically-limited service. IEEE J. Sel. Areas Commun. **9**(2), 185–193 (1991)
20. Wolff, R.: Poisson arrivals see time averages. Oper. Res. **30**(2), 223–231 (1982)
21. van Ommeren, J.C.W.: The discrete-time single-server queueing model. Queueing Syst. **8**(1), 279–294 (1991)
22. Ross, S.M.: Stochastic Processes. Wiley, New York (1996)
23. Keilson, J., Servi, L.D.: A distributional form of Little's law. Oper. Res. Lett. **7**(5), 223–227 (1988)
24. Fuhrmann, S.W., Cooper, R.B.: Stochastic decompositions in the M/G/1 queue with generalized vacations. Oper. Res. **33**(5), 1117–1129 (1985)
25. Feller, W.: An Introduction to Probability Theory and its Applications, vol. II. Wiley, New York (1966)